Finite mixture models in the tails

Marilena Furno, Francesco Caracciolo

Department of Agricultural Sciences, Università degli Studi di Napoli "Federico II"

Abstract

The finite mixture model allows to estimate regression coefficients for different groups of the data under investigation determining endogenously the number of groups in which the observations are segmented. In what follows the analysis is extended beyond the mean, estimating the model in the tails of the conditional distribution of the dependent variable within each group. While the clustering reduces the overall heterogeneity, since the model is estimated for similar observations, the analysis in the tails uncovers within groups heterogeneity and/or skewness. A Monte Carlo experiment and two empirical applications conclude the analysis. In the well-known birthweight data set the finite mixture model identifies and computes the regression coefficients of different groups, each one with its own characteristics, both at the mean and in the tails. In the family expenditure data, the analysis of within and between groups heterogeneity provides interesting economic insights on price elasticities. The analysis in classes proves to be more efficient than the model estimated without clustering. By extending the finite mixture approach to the tails consents a more accurate investigation of the data, unveiling sources of heterogeneity and asymmetry otherwise left undetected.

Keywords: finite mixture model, expectation-maximization algorithm, quantiles, expectiles

1. Introduction

The finite mixtures (FM) approach (McLachlan and Peel 2004) identifies homogeneous groups within the sample, meanwhile computing the coefficients of the regression model within each group. The data are aggregated into clusters sharing functionally similar patterns, and the clustering process is simultaneous to the estimation process. This allows us to compute class-specific coefficients without any prior knowledge or assumption on the clustering, meanwhile accounting for group heterogeneity. Finite mixtures are indeed commonly employed as a flexible modeling device to account for heteroskedasticity (Compiani and Kitamura, 2016).

Due to data heterogeneity, the sample may form several distinct clusters, and FM adaptively models multiple regressions, each one responsible for one subset/cluster of the data. These models have been widely used in market segmentation studies, or patients' disease progression subtyping, (Städler et al. (2010), Khalili (2011), Khalili and Chen (2007)). Liang et al (2018) extend the FM approach to improve robustness. A market segmentation approach is implemented by Betrail and Callavet (2008) for fruit and vegetable consumption, by Caudill and Mixon (2016) and by Caracciolo and Furno (2020) for wine purchases. The latter define segments of consumers meanwhile estimating the model within each class, and in a second step investigates hedonic functions in the tails, away from the conditional mean, by means of quantile estimators. By computing FM in the tails, namely FMQ, it is possible to evaluate the presence of within-class heterogeneity in addition to the dispersion among groups explained by the clustering. The quantile regression estimator introduces an asymmetric weighting system to move the estimated line away from the conditional mean. Once the classes are defined by the FM approach, we can robustly estimate within each group the coefficients at different points of the conditional distribution. Alfò et al. (2017) present a finite mixture of quantile and M-quantile regression to model random coefficients. The random effects represent individual-specific unobserved heterogeneity, while observed covariates describe the observed counterpart. This model, although using the FM approach within the quantile regression framework, focuses on random effect

coefficients and substantially differs from the model here considered. Another related approach to FM analyzes the joint density p(y, x) for the dependence of y on x within clusters, defining the cluster weighted model, CWM (Gershenfeld 1997, Gershenfeld et al. 1999). The joint distribution of dependent and explanatory variables is decomposed as the product of the marginal and the conditional distributions. The marginal is a density for values of x, while the conditional models y/x. The relative sizes of these functions between the clusters determines whether an observation is associated with any given cluster. Thus, the joint density becomes a mixture of two distributions, x and y/x, and they are generally assumed to be normal. To improve robustness of CWM, Lin et al. (2007) consider a mixture of skewed Student-t distributions, while Garcia-Escudero et al. (2017) assume normality meanwhile adding weights to trim outliers.

Summarizing, FMQ merges the advantages of routine clustering of the FM models, that provides a more detailed group-specific analysis, with the benefits of quantiles, that analyze the behavior of each group beyond the mean, in the tails of the conditional distribution. The combination of clustering and tails analysis yields a deeper interpretation of the data. Without FM, segmentation cannot be assessed, and by implementing OLS instead of quantiles within-group heterogeneity and asymmetry cannot be captured.

In what follows a Monte Carlo experiment analyzes the behavior of FMQ and two empirical applications estimate multi-class models: birthweight and households' expenditures, two well-known data sets, are examined to identify distinct classes in each sample, each class with its own estimated coefficients related to individual differences. The results show that the combination of clustering and quantiles in FMQ provides a better explanation and turns out to be more efficient than the model estimated without clustering.

2. The empirical approach

The FM approach explicitly considers possible sources of heterogeneity, consenting to identify homogeneous classes without any prior knowledge of the number of clusters. FM is a cluster-wise regression analysis, capable of identifying observations that exhibit similar behavior (Melnykov and Maitra, 2010). Unlike more traditional approaches, there are no prior assumptions about the clusters. The definition of the groups is simultaneous with the estimation of the regression parameters and the estimation procedure is based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The conventional approaches instead compute in a first stage the clusters and in a second stage estimate the regression coefficients within each class.

In what follows the dependent variable y_i is function $f(x_i\ \beta)$ of a set of $q + p$ coefficients related to the row vector for the ith observation $x_i = (x_{i1}\ x_{i2})$, in a sample of size n. The maximum likelihood objective function, which provides the maximization step, relates y to x_2 and is given by

$$\prod_n \pi_{ij}\, f(y_i | x_{i2}, z_i = 1) \qquad (1)$$

where π_{ij} is the probability of observation i=1, …, n to belong to a given group j=1, …, K, K is the number of endogenously determined groups in the model, and z_i is a non-observable latent variable determining for each observation its involvement in one group or another. The likelihood combines the conditional likelihood of each latent class weighted by the associated latent-class probability. For each observation, the probability of inclusion in group j, π_{ij}, is function of z_i, which assumes values 0 or 1 to define the exclusion or the inclusion of the ith observation in the specific j^{th} group

$$\pi_{ij} = \pi(z_i\text{=}1) = \exp(z_i)/\Sigma_{j=1,\,K} \exp(z_j) \qquad (2)$$

Equation (2) yields the expectation step of the algorithm.

Unfortunately, we do not know from which group an observation is drawn, the observed sample is incomplete since z_i is an unobservable latent variable. However, z_i can be approximated by a set of explanatory variables, x_{1i}. A multinomial logit model computes π_{ij} as follows

$$\pi_{ij} = h(x_{i1}) = \exp\,(g(x_{i1}))/\Sigma_{j=1,K} \exp(g(x_{j1})) \qquad (3)$$

where $g(x_{i1})$ is the function relating the probability of being in the j class to the x_{i1} characteristics.

Next, the main model explains the dependent variable y_i as a function of the variables in x_{i2}. The idea is that each observation has its own characteristics $f_j(x_{i2}\, \beta_j)$, and its own probability of belonging to a group $h_j(x_{i1})$. The contemporaneous estimation of the probability to belong to a group π_{ij} together with the p regression coefficients β_j in each of the K groups is a difficult task, at times intractable. Therefore, the EM algorithm breaks the objective function iterating between the estimated expected probability of belonging to a specific group, $E(\pi_{ij})$ in (3), and the estimates of the p regression coefficients for each group $\widehat{\beta}_j$ in (1). This estimation approach allows us to model bimodal/multimodal distributions and more in general finite mixtures of distributions. The bimodality/multimodality of the y distribution allows us to consider the presence of two/many distinct groups.

In detail, starting from initial prior probabilities, for instance $\pi_j^0 = 1/K$ for an arbitrarily selected K, at iteration s the expectation step computes the posterior probabilities as

$$\pi_j^s = E(\pi_{ij}) = \pi_j^{s-1} f(x_{i2}\beta_j^{s-1}) \Big/ \textstyle\sum_{j=1,K} \pi_j^{s-1} f(x_{i2}\, \beta_j^{s-1}) \tag{4}$$

and by modeling K separate distributions, one for each group, the maximizing step computes $\sum_{i=1,n} \pi_j^s f(x_{i2}\, \beta_j)$, yielding an updated estimate of the p-vector $\hat{\beta}_j^s$. The iterations continue until convergence yielding, after m iterations, the final $\hat{\pi}_j^m$ and $\widehat{\beta}_j^m$. The $\hat{\pi}_j^m$, that provide for each observation the membership probability to each class, are included in equation 1 to allocate each observation to its group thus yielding the estimates of $f(x_{i2}\, \beta_j)$ within each class.

3. Finite Mixture in the tails

Once the model in each class is estimated, the focus moves to its behavior in the tails, away from the conditional mean. This analysis uncovers heterogeneity and/or asymmetric patterns within each group. The key element used to investigate the within-group behavior is provided by the final $\hat{\pi}_j^m$ weights of

equation (4). Their role is to select the observations belonging to group j. Within each group, quantile regressions (Koenker and Basset, 1978; Koenker, 2005) allow us to estimate a regression in the tails, measuring the impact of the explanatory variables at different levels of the dependent variable: not only at the center, the mean or the median, but also in the tails of the distribution, such as the first and third quartiles of the conditional distribution. This is particularly informative in the case of asymmetry or heteroskedasticity.

Assume a linear regression model $y_{ij}=x_{i2}'\beta_j + e_{ij}$ within each class. The final step separately implements quantile regressions (Koenker, 2005) within each group. The quantile regression estimator maximizes the following objective function

$$\sum_i \{\tau - 1(y_i \leq x_{i2}\beta_j\ )\}|y_i - x_{i2}\beta_j| \tag{5}$$

within each group. In equation (5) the absolute value of the regression errors, $|y_i - x_{i2}\beta_j|$, are asymmetrically weighted by τ or (1-τ) to compute the selected quantile. These weights set the position of the estimated line away from the center of the conditional distribution. For instance, the 25th quantile regression is computed by assigning weights τ = 0.25 to the positive residuals and (1- τ) = 0.75 to the negative ones, thus attracting toward the first quartile the estimated regression. The quantile regression estimator provides results that are robust to anomalous values, unless the model is estimated at extremely low/high quantiles. Within the FM framework equation (5) becomes

$$\sum_i \{\tau - 1(y_{ij} \leq x_{i2}\beta_j\ )\}\pi_j^m \left|y_{ij} - x_{i2}\beta_j\right| \tag{6}$$

The $\hat{\pi}_j^m$ weights, which define the probability of belonging to a given class after m iterations, provide class-specific quantile regression estimates. This is a two-step approach: first, the EM algorithm is implemented to compute the probability weights defining the various groups, and then the analysis is extended to the tails by computing the quantile regressions within each class. An analysis away from the conditional mean within each group is implemented to verify the presence of within-group heterogeneity and/or to look for within-group asymmetry. The EM approach provides the tool to cluster

the data set by means of the posterior probabilities of each observation to belong to a given class. Multiplying the observations by their probability of belonging to class 1, for instance, and computing equation (6) at different quantiles allows us to find if the impact of the explanatory variables on *y* is constant across quantiles or changes according to the selected quantile within class 1. The same steps are repeated for class 2 and for all the estimated groups. Computing equation (6) within each class at different quantiles implements the FM estimator at the quantiles, namely the FMQ estimator. FMQ uncovers within group heteroskedasticity and/or skewness.

4. Monte Carlo study

To check the properties of the FMQ estimator at various quantiles, a Monte Carlo experiment is considered. The model is estimated at the first, second and third quartile. The tables report the sample mean, the standard deviation of the true coefficients and of the estimates implementing equations (4) and (5) in 5,000 iterations. The standard OLS and quantile regression estimator are also computed in the simulations.

In each iteration, the outcome $\mathbf{Y} = \pi f(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta})$ is generated in a sample of 3,000 observations as a mixture of two separate distributions, Y_1 and Y_2, given by:

$$Y_{i1} = -0.90 + 0.70X_i + 0.2W_i + 0.6W_iX_i + \eta_i$$

$$Y_{i2} = -0.45 + 0.35X_i + 0.1W_i + 0.3W_iX_i + \varepsilon_i \qquad (7)$$

where the explanatory variables **X** and **W** are defined as: **X** is drawn from a standard normal distribution, $\mathbf{X} \sim N(0, 1)$, and **W** is a dummy variable assuming values 0 and 1 each with probability 0.5, $\mathbf{W} \sim B(0.5)$. The error terms are i.ni.d,: independent one another and, to generate heteroskedasticity in the mixture **Y**, the error term in Y_{i1}, has variance greater than in Y_{i2}: $\eta_i \sim N(0, 0.5)$; $\varepsilon_i \sim N(0, 0.25)$. The changing variance between the two distributions causes heteroskedasticity in $\mathbf{Y}_i$, which determines a change in the $\hat{\boldsymbol{\beta}}$ parameters estimated across quantiles.

The mixture distribution of **Y** is defined by the following link function:

$$\mathbf{Y}_i = \begin{cases} Y_{i1}, & \pi_i < \mu(\pi) \\ Y_{i2}, & \pi_i \geq \mu(\pi) \end{cases} \tag{8}$$

with $\mu(\pi)$ being the sample mean of π_i, and the latter is defined as:

$$\pi_i = \frac{1}{1 + \exp(-0.6 + 0.9Z_i + 0.3D_i)} + u_i \tag{9}$$

Z is drawn from a standard normal distribution, **Z** ~ N(0, 1); **D** is a dummy variable assuming values 0 and 1with probability 0.5, **D** ~B(0.5); the error term is normally distributed as **u**~ N(0, 0.5).

5. Birthweight Data

The following case study considers a well-known dataset analyzing the impact of demographic characteristics and maternal behavior on birthweight of live infants born in the USA in June 1997, considering mothers of age comprised between 18 and 45, in a sample of size n=198,377 (Detailed Natality Data, National Center for Health Statistics). Previous studies by Chernozhukov and Fernandez-Val (2011), Abrevaya (2001), Koenker and Hallock (2001) use the same dataset but analyze different issues. The data include the outcome of interest, birthweight, and covariates comprising socio-economic characteristics, the *q*-row vector for the ith observation of x_{i1}, to describe mother conditions that may influence the outcome: ethnicity, marital status, mother age and age squared, education grouped in three classes (high school, some college education, completed college). These variables determine the probability to belong to a group. Next mother conditions during pregnancy are gathered in x_{i2}: smokers, number of cigarettes smoked per day, weight gained during pregnancy, baby gender, trimester of prenatal visits - mothers having prenatal medical care in the first, in the second, in the third trimester of pregnancy, or having no prenatal medical care at all. These variables provide the regressors in x_{2i} and define pregnancy conditions. Mothers having prenatal care in the first trimester comprise 84% of the sample. The other two groups cover, respectively, 13% and 2% of the sample. Mothers having no prenatal care at all are about 0.8% of the sample. The idea is that each woman has her own

characteristics, $f_j(x_{i2}\ \beta_j)$, and each baby has its own probability of belonging to a group, $g(x_{i1})$. Table 1 reports the main descriptive statistics of the sample.

Table 2 reports the OLS and the quartile regression estimates in the entire sample without any clustering. The model without any grouping provides a good term of comparison. In this table all the variables are statistically significant, OLS and quantile regressions yield coherent results, but the latter estimator provides changing coefficients across quartiles, thus signaling heteroskedasticity and/or skewness. In detail, black has a negative impact on birthweight, particularly at the lower quartile; college education and marital status have the widest impact at low birthweight together with the timing of prenatal visit. Smoke, cigarettes per day, mother weight gain and age do not change much across quartiles, while baby gender has a greater impact at high weight.

Next, we implement the standard FM approach. Table 3 reports summary statistics when selecting different numbers of group. In the table the improvement in the loglikelihood, the BIC and the Akaike criteria is sizable when comparing the 1-group with the 2-groups model and the latter with the 3-groups clustering. Vice versa, the improvement granted by the 4-groups clustering is less relevant. Since, in addition, the higher number of groups makes more cumbersome the comparison of the results, a 3-groups clustering is selected.

Table 4 reports the estimated impact of the x_{i1} variables to define the probability to belong to group 2 and 3 with respect to the first group. Marital status, education and ethnicity are the determinants of the probabilities to belong to each group. Group 1 gathers unmarried and black women, group 3 collects non-black educated married mothers, while in group 2 some college education has a negative impact and is the sole statistically significant explanatory variable.

Table 4 considers the impact of the x_{i2} variables on birthweight within each group and across quartiles, thus providing the FMQ results. In the first group, gathering unmarried and black women, smoke has the widest negative impact on birthweight, the largest impact of mother weight gain, the largest medicare coefficients and the smallest impact of the gender coefficient. Across quartiles the

estimates do change: the relevance of medicare, mother age and smoke decreases as birthweight grows. This reveals the presence of heteroskedasticity or skeweness within this group.

In the second group, collecting mothers with some college education, the estimates are smaller than in the first group: smoke shows a less relevant negative impact; medicare has a greater impact at the first trimester but is everywhere smaller than in the first group; mother weight gain impact is very small as well with respect to the first group; vice versa, mother age coefficients are comparable to those estimated for the first group. In group 2 the change in the estimated coefficients across quartiles is less sizable, showing a greater within group homogeneity. Finally group 3, bundling educated and married white mothers, shows an almost constant impact of smoke, mother age and weight gain across quartiles, a small and declining impact of medicare, and the largest impact of baby gender, that is almost constant across quartiles.

Figure 1 depicts the probability to belong to each group. The probability distribution of the second and the third group is quite spread out, showing some concentration on low values for the second group and on higher values for the third one. In the first group the probability is highly concentrated to the left, at low values.

Figure 2 displays the empirical distributions within each group and across quartiles. The rows display the quartiles while the columns depict the groups. The graphs in the first row provide the empirical distributions within each group at the first quartile, at the median in the second row and at the top quartile in the last row. The graphs in the first column show the great skewness of the empirical distributions of the first group at all quartiles, all showing a long right tail. The second column, depicting the empirical distributions of group 2, presents graphs more symmetrical and closer to normality, with slightly longer right tail. The right column, collecting group 3 graphs, shows homogeneity as well as closeness to normality, but with a slightly longer left tail.

A final consideration can be drawn by comparing standard quantile regression and FMQ results, focusing on the variables in common. For instance, the trimester variables have a declining impact in

table 2, while in table 5 for the third group they assume a quite small and almost constant value at all quartiles. Smoke declines across quartile in table 2, while it is constant in table 5 for group 3. The clustering offers a completer and more detailed picture of the model.

6. Family expenditure data

In what it follows, FMQ has been implemented to analyze a naive demand function using the Canadian Family Expenditure Survey data in Lewbel and Pendakur (2009), Pendakur et al., (2010) and Pendakur and Sperlich (2010). While the original sample comprises several items[1], in this example only family food consumption and relative prices are taken into consideration. The sample is based on 4,847 observations on food prices and quantities consumed in the period 1969 - 1996.

The food quantities consumed at home (in logarithm) represent the dependent variable. Recalling that in FMQ each observation has its own characteristics $f_j(x_{i2}\ \beta_j)$ and its own probability of belonging to a group $h_j(x_{i1})$, we have two different set of covariates: x_{i2} including normal factors influencing demand like at-home and outside-home food prices (in logarithm), total expenditures of the family (in logarithm), and a time variable equal to the calendar year minus 1986; x_{i1} are the variables determining the probability to belong to a group and include socio-economic characteristics of the family like age minus 40 of the household head; a gender dummy (one for men); a car non-owner dummy and the share of family expenditures in outside-home food. Table 6 reports sample mean and standard deviation for the variables of interest. Following the same procedure described in the previous example, a three-classes model has been selected according the model selection criteria (AIC and BIC). Table 7 reports the estimated impact of the x_{i1} variables to define the probability to belong to group 1 and 3 with respect to the second group. Compared to the other two groups, group 1 gathers families with an older female household head, without car ownership and with a lower percentage of expenditure in out of home

[1] Original dataset includes nine categories: food-in, food-out, rent, clothing, household operation, household furnishing and equipment, transportation operation, recreation, and personal care (Lewbel and Pendakur, 2009).

food. Group 3 is similar to group 2 with the sole difference of a lower percentage of expenditure in out of home food. Thus, group 2 includes mainly families with household head male, younger, with the highest percentage of expenditure in food out of home and car owner.

Table 8 estimates the influence on food demand of the x_{i2} variables within each group and across quartiles, providing the FMQ results at the median. Since both dependent and independent variables are log values, the estimates coefficients can be naively interpreted as demand elasticities.

Estimates from the first group indicate that older female household head, without car and with a lower percentage of expenditure in out of home food present the most anelastic food price elasticity, and the lowest expenditure elasticities, while the second group (younger, male household head, with car and with the highest percentage of expenditure in out of home food) shows the more elastic food price elasticity and the highest expenditure elasticity. Since group 3 and group 2 differ only in the share of expenditure in food outside of home, with the latter group characterized by an higher percentage of the expenditure share in out of home food consumption, it is interesting to see how the elasticities may differ among the two groups. Indeed, food out of home is considered as complement for group 3 families (-0.227), while it is rather a substitute for group 2 (0.771). This result reveals the validity of exploring the sample heterogeneity through FMQ.

Finally, the table reports the interquartile regression results comparing 25th and 75th quantiles. They show that only group 1 has statistically significant estimates of the difference between upper and lower quartile coefficients. In detail p_1, the price elasticity of at-home food, decreases at the upper tail, while income and time decrease. The FMQ estimates at the quantiles show that most of the volatility of the sample has been abated by the grouping mechanism. However, there is some left-over volatility in group 1. This effect could not be detected by the standard FM approach nor, perhaps, by a standard heteroskedasticity test. The presence of volatility in only one out of three groups could easily be undetected. Indeed, the Koenker test of heteroskedasticity - using the squared OLS residuals in an auxiliary regression with squared independent variables as regressors, $ln(p_1)^2$, $ln(p_2)^2$, $ln(Y)^2$ - provides

no statistically significant coefficient but the constant term and does not reject the null of homoskedasticity.[2]

7. Conclusions

The finite mixtures (FM) approach identifies homogeneous groups within the sample, meanwhile computing the coefficients of the regression model in each group. The data are aggregated into clusters sharing functionally similar patterns, and the clustering process is simultaneous to the estimation process. This allows us to compute class-specific coefficients without any prior knowledge or assumption on the clustering, meanwhile accounting for group heterogeneity. Finite mixtures are indeed commonly employed as a flexible modeling device to account for heteroskedasticity. The FM approach is usually implemented at the average value of the conditional distribution, and here the same investigation is extended to the tails, at the quantiles (FMQ). Once estimated the probability to belong to a group, the analysis moves to investigate the cluster-wise regressions along the entire conditional distribution, at the median and in the tails. A Monte Carlo experiment is here implemented to verify the behavior of the Finite Mixture at the quantiles, FMQ.

Future research will attempt to further generalize FMQ. Among the possible issue is, for instance, to allow the clusters to change across quantiles. In this study the clusters are constant while the regression coefficients can change across quantiles due to skewness and/or heteroskedasticity. It would be possible to consider the probability to belong to a given class as non-constant, changing across quantiles. For instance, in the case study on birthweight here analyzed, the clustering is function of mothers' socio-economic characteristics. Instead of looking at the average impact of characteristics on

[2] The estimates in the auxiliary regression are $ln(p_1)^2$= .266 with t=0.82, $ln(p_2)^2$= -.376 with t=-1.13, $ln(Y)^2$=-.027 with t=-0.59, intercept=.335 with t=10.9. Only the intercept is statistically significant, thus showing that the variance is constant in the sample.

the probability of belong to a given class, the question becomes how the clustering would change with respect to very good or very poor socio-economic characteristics.

References

Abrevaya J., Dahl C., 2001. The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. Empirical Economics 26, 247–257.

Ahmadi-Esfahani, F. Z. and Stanmore, R. G. 1997. Export demand for attributes of Australian wheat in Asia and the Middle East. Food Policy 22(2), 145-54.

Alfò M., Salvati N., Ranalli M. G., 2016. Finite mixtures of quantile and M-quantile regression models. Statistics and Computing 27, 547–570.

Betrail P., Callavet F., 2008. Fruit and vegetable consumption: a segmentation approach. American Journal of Agricultural Economics 90, 827-842.

Caudill, S. B., Mixon Jr F. G., 2016. Estimating class-specific parametric models using finite mixtures: an application to a hedonic model of wine prices. Journal of Applied Statistics 43(7), 1253-1261.

Caracciolo F., Furno M., 2020. Hedonic Functions, Heterogeneous Consumers, and Wine Market Segmentation, Journal of Agricultural and Resource Economics, 45(2):300–317.

Chernozhukov V., Fernandez-Val I., 2009. Inference for Extremal Conditional Quantile Models, with an Application to Market and Birthweight Risks. Review of Economic Studies 78, 1-42.

Combris P., Lecocq S. and Visser M., 1997. Estimation of a hedonic price equation for bordeaux wine: does quality matter? The Economic Journal 107(441), 390-402.

Compiani C., Kitamura Y., 2016. Using mixtures in econometric models: a brief review and some new results. The Econometrics Journal 19, C95–C127.

Dempster P., Laird N., Rubin D., 1977. Maximum likelihood for incomplete data via the EM algorithm. Journal of the Royal Statistical Society (B) 39, 1-38.

García-Escudero L. A., Gordaliza A., Greselin F., Ingrassia S., Mayo-Iscar A., 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. Statistics and Computing 27, 377-402.

Gershenfeld, N., 1997. Nonlinear inference and cluster-weighted modeling. Annals New York Academy of Science 808(1), 18-24 (1997)

Gershenfeld, N., Schoner, B., Metois, E., 1999. Cluster-weighted modelling for time-series analysis. Nature 397, 329-332.

Khalili A., 2010. New estimation and feature selection methods in mixture-of-experts models. The Canadian Journal of Statistics 38, 519-539.

Khalili A., Chen J., 2007. Variable Selection in Finite Mixture of Regression Models, Journal of the American Statistical Association 102, 1025-1038.

Koenker R., 2005. Quantile regression, Cambridge University Press.

Koenker R., Allock K., 2001. Quantile Regression: An Introduction. Journal of Economic Perspectives 15, 143-156.

Koenker R., Bassett G., 1978. Regression quantiles. Econometrica 46, 33–50.

Lewbel, A. and K. Pendakur. 2009. Tricks with Hicks: The EASI Demand System. American Economic Review 99: 827–863.

Liang J., Chen K., Lin M., Zhang C., Wang F., 2018. Robust finite mixture regression for heterogeneous targets. Data Mining and Knowledge Discovery 32, 1509–1560.

Lin T. I., Lee J. C., Hsieh W. J., 2007. Robust mixture modeling using the skew t distribution. Statistics and Computing 17, 81–92.

Melnykov V., Maitra R, 2010, Finite mixture models and model-based clustering. Statistics Surveys 4, 80-116.

McLachlan G., Peel D., 2004. Finite Mixture Models, Wiley Series in Probability and Statistics.

Pendakur, K., Scholz, M., & Sperlich, S. (2010). Semiparametric indirect utility and consumer demand. Computational statistics & data analysis, 54(11), 2763-2775.

Pendakur, K., & Sperlich, S. (2010). Semiparametric estimation of consumer demand systems in real expenditure. Journal of Applied Econometrics, 25(3), 420-457.

Städler N., Bühlmann P., van de Geer S., (2010). $\ell 1$-penalization for mixture regression models. Test 19, 209–256.

Table 1: Birthweight variables, summary statistics

	Mean	Std. Dev.
birthweight	3.369	0.570
black	0.161	0.367
married	0.713	0.452
boy	0.514	0.500
trimester1	0.843	0.364
trimester2	0.126	0.332
trimester3	0.022	0.147
novisit	0.008	0.090
education_highschool	0.349	0.477
education_somecollege	0.242	0.428
education_college	0.249	0.433
mom_age	27.422	5.705
smoke	0.132	0.338
cigsper	1.505	4.722
mom_weightgain	30.699	12.902
mom_age2	784.558	323.650

Note. observations 198,377

Table 2: Regression estimates at the quartiles without clusters

	.25			.50			.75			OLS		
	Coef.	Std. Err.	t	Coef.	Std. Err.	t	Coef.	Std. Err.	t	Coef.	Std. Err.	t
black	-.1940	.0043	-44.40	-.1736	.0037	-46.12	-.1677	.0041	-40.28	-.1997	.0035	-55.89
married	.0581	.0039	14.62	.0521	.0033	15.23	.0436	.0037	11.52	.0606	.0032	18.66
educ_highsch	.0193	.0045	4.21	.0127	.0039	3.23	.0181	.0043	4.15	.0154	.0037	4.12
educ_smcol	.0381	.0051	7.44	.0293	.0044	6.65	.0312	.0048	6.41	.0318	.0041	7.60
educ_college	.0530	.0055	9.65	.0333	.0047	7.04	.0259	.0052	4.95	.0379	.0044	8.44
mom_age	.0334	.0024	13.71	.0317	.0021	15.11	.0347	.0023	14.94	.0363	.0019	18.23
smoke	-.1529	.0076	-19.89	-.1677	.0066	-25.31	-.1665	.0073	-22.73	-.1681	.0062	-26.75
cigsperday	-.0048	.0005	-8.80	-.0036	.0004	-7.66	-.0034	.0005	-6.55	-.0037	.0004	-8.31
mom_age2	-.0005	.0000	-12.36	-.0004	.0000	-12.47	-.0004	.0000	-11.84	-.0005	.0000	-15.83
mom_wtgain	.0083	.0001	72.67	.0075	.0000	76.04	.0070	.0001	64.45	.0089	.0001	94.26
boy	.0967	.0029	32.71	.1190	.0025	46.71	.1329	.0028	47.18	.1088	.0024	45.02
trimester1	.1936	.0166	11.65	.1425	.0143	9.96	.1326	.0158	8.38	.1848	.0135	13.61
trimester2	.1997	.0169	11.77	.1422	.0146	9.72	.1366	.0161	8.44	.1966	.0138	14.16
trimester3	.2121	.0192	11.05	.1441	.0165	8.71	.1230	.0182	6.72	.2146	.0157	13.67
constant	2.057	.0373	55.13	2.440	.0321	75.92	2. 710	.0355	76.25	2.280	.0305	74.77
pseudo R^2	0.0591			0.0584			0.0568			0.1101		
observations	198377			198377			198377			198377		

Table 3: Akaike's information criterion and Bayesian information criterion to select the groups

	n	loglikelihood	AIC	BIC
1 group	198377	-158514.2	317058	317211
2 groups	198377	-149399.2	298854	299140
3 groups	198377	-147770.5	295631	296089
4 groups	198377	-147417.6	294959	295591

Table 4: finite mixture model in the tails, estimated probability, n= 198377

	Coef.	Std. Err.	z	p-value
Class 2				
black	-*.0313*	.0845	-0.37	0.71
married	-*.0461*	.0585	-0.79	0.43
educ_highsch	-*.0815*	.0512	-1.59	0.11
educ_smcol	-.2019	.0644	-3.10	0.00
educ_college	-*.0170*	.0829	-0.21	0.83
constant	1.4214	.0787	18.06	0.00
Class 3				
black	-1.853	.0704	-26.31	0.00
married	.61515	.0424	14.48	0.00
educ_highsch	.11582	.0462	2.50	0.01
educ_smcol	.24704	.0518	4.77	0.00
educ_college	.42802	.0573	7.46	0.00
constant	1.6548	.0744	22.23	0.00

Table 5: Finite mixture at the quartiles

Class 1	.25 Coef.	Std. Err.	z	p	.50 Coef.	Std. Err.	z	p	.75 Coef.	Std. Err.	z	p
smoke	-.1832	.0033	-55.45	0.00	-.1757	.0029	-60.06	0.00	-.1126	.0071	-15.67	0.00
tri1	.43743	.0071	61.57	0.00	.31928	.0053	59.59	0.00	.20840	.0131	15.83	0.00
tri2	.58677	.0074	78.56	0.00	.38788	.0055	70.18	0.00	.34778	.0135	25.61	0.00
tri3	.85633	.0101	84.76	0.00	.55053	.0065	84.88	0.00	.54848	.0159	34.42	0.00
cigsper	.00073	.0002	2.99	0.00	-.0004	.0002	-1.87	0.06	-.0045	.0005	-8.52	0.00
mom_wtgain	.02032	.0000	409.1	0.00	.02002	.0000	455.4	0.00	.02256	.0001	207.11	0.00
boy	.06218	.0013	45.92	0.00	.06145	.0011	51.28	0.00	.05748	.0029	19.53	0.00
mom_age	.06468	.0010	62.29	0.00	.04256	.0009	46.31	0.00	.01666	.0022	7.37	0.00
mom_age2	-.0012	.0000	-65.98	0.00	-.0008	.0000	-50.59	0.00	-.0005	.0000	-12.55	0.00
constant	.62905	.0152	41.15	0.00	1.2772	.0135	94.38	0.00	.20509	.0007	268.78	0.00
Class 2	**Coef.**	**Std. Err.**	**z**	**p**	**Coef.**	**Std. Err.**	**z**	**p**	**Coef.**	**Std. Err.**	**z**	**p**
smoke	-.1383	.0056	-24.48	0.00	-.1210	.0029	-40.74	0.00	-.0904	.0031	-28.63	0.00
tri1	.23220	.0108	21.38	0.00	.21458	.0057	37.58	0.00	.16534	.0060	27.25	0.00
tri2	.19792	.0111	17.74	0.00	.18557	.0058	31.64	0.00	.16508	.0062	26.49	0.00
tri3	.19135	.0126	15.07	0.00	.18780	.0066	28.14	0.00	.16793	.0071	23.68	0.00
cigsper	-.0045	.0004	-11.12	0.00	-.0044	.0002	-20.45	0.00	-.0053	.0002	-23.14	0.00
mom_wtgain	.00746	.0000	85.51	0.00	.00717	.0000	156.17	0.00	.00703	.0000	144.24	0.00
boy	.08215	.0023	35.59	0.00	.07748	.0012	63.86	0.00	.06308	.0012	48.92	0.00
mom_age	.06356	.0017	35.75	0.00	.05009	.0009	53.58	0.00	.03641	.0009	36.65	0.00
mom_age2	-.0010	.0000	-33.41	0.00	-.0008	.0000	-51.88	0.00	-.0007	.0000	-39.92	0.00
constant	1.6658	.0263	63.13	0.00	2.1023	.0138	151.54	0.00	2.5629	.0147	173.85	0.00
Class 3	**Coef.**	**Std. Err.**	**z**	**p**	**Coef.**	**Std. Err.**	**z**	**p**	**Coef.**	**Std. Err.**	**z**	**p**
smoke	-.1712	.0051	-33.47	0.00	-.1774	.0063	-28.19	0.00	-.16460	.0062	-26.53	0.00
tri1	.07430	.0126	5.87	0.00	.03108	.0155	2.00	0.04	.05174	.0153	3.37	0.00
tri2	.07563	.0129	5.85	0.00	.04788	.0158	3.01	0.00	.06578	.0156	4.20	0.00
tri3	.04907	.0144	3.40	0.00	.03427	.0177	1.93	0.05	.03830	.0175	2.19	0.02
cigsper	-.0009	.0003	-2.70	0.00	-.0039	.0004	-8.99	0.00	-.00433	.0004	-9.96	0.00
mom_wtgain	.00643	.0000	84.56	0.00	.00625	.0001	66.84	0.00	.00626	.0001	67.95	0.00
boy	.14252	.0019	74.72	0.00	.13394	.0023	57.10	0.00	.14247	.0023	61.60	0.00
mom_age	.02695	.0015	17.57	0.00	.02493	.0018	13.22	0.00	.02838	.0018	15.25	0.00
mom_age2	-.0003	.0000	-11.75	0.00	-.0003	.0000	-9.41	0.00	-.00036	.0000	-11.14	0.00
constant	2.4562	.0248	98.78	0.00	2.7955	.0305	91.43	0.00	2.9825	.0301	98.92	0.00

Table 6: Family expenditure data, summary statistics

Variable	Description	Mean	Std.dev
ln(q)	Logarithm of quantity consumed of in home food	-2.147	0.586
ln(p_1)	Logarithm of in home food price	-0.048	0.432
ln(p_2)	Logarithm of out home food price	0.038	0.487
ln(Y)	Logarithm of Total expenditure	-0.112	0.590
Time	Time variable (year minus 1986)	3.036	8.703
Age	Age minus 40	0.701	11.873
Gender	Sex (1 if male)	0.510	0.500
Car	Car non-owner (1 if non-owner)	0.413	0.492
Ohe	Percentage out of home food expenditure	7.358	6.984

Table 7: finite mixture model in the tails, estimated probability, n= 4,840

	Coef	Std.error	z-stat	p-value
Class1				
Age	0.053	0.014	3.860	0
Gender	-1.160	0.327	-3.550	0
Car	0.877	0.305	2.880	0.004
O.H.Perc.	-0.238	0.021	-11.310	0
Cons	5.158	0.399	12.940	0
Class 3				
Age	-0.019	0.014	-1.410	0.16
Gender	0.123	0.316	0.390	0.697
Car	-0.133	0.296	-0.450	0.652
O.H.Perc.	-0.111	0.016	-6.990	0
Cons	4.077	0.387	10.520	0

Table 8: Finite mixture at the median and interquartiles (75th - 25th) estimates

	τ = .50				Interquartile			
	Coef	Std.error	z-stat	p-value	Coef	Std.error	z-stat	p-value
Class 1								
ln(p_1)	-0.574	0.079	-7.31	0	-0.276	0.143	-1.93	0.053
ln(p_2)	-0.009	0.088	-0.1	0.918	0.173	0.154	1.13	0.259
ln(Y)	0.512	0.01	51.82	0	0.034	0.017	1.94	0.052
Time	-0.005	0.002	-2.59	0.01	0.007	0.003	2.1	0.036
Const	-1.804	0.009	-207.04	0	0.388	0.012	31.6	0
Class 2								
ln(p_1)	-1.928	0.074	-25.92	0	1.736	1.551	1.12	0.263
ln(p_2)	0.771	0.079	9.76	0	-2.075	1.327	-1.56	0.118
ln(Y)	0.769	0.008	97.14	0	0.241	0.226	1.07	0.285
Time	0.008	0.002	4.79	0	0.041	0.025	1.63	0.102
Const	-3.562	0.006	-592.3	0	0.955	0.211	4.53	0
Class 3								
ln(p_1)	-0.614	0.093	-6.63	0	-0.002	0.180	-0.01	0.99
ln(p_2)	-0.227	0.106	-2.15	0.031	0.161	0.212	0.76	0.448
ln(Y)	0.606	0.013	45.13	0	0.127	0.031	4.07	0
Time	0.001	0.002	0.5	0.619	-0.004	0.005	-0.79	0.428
Const	-2.317	0.008	-280.05	0	0.495	0.016	31.02	0

Figure 1: Probability to belong to each class at the mean

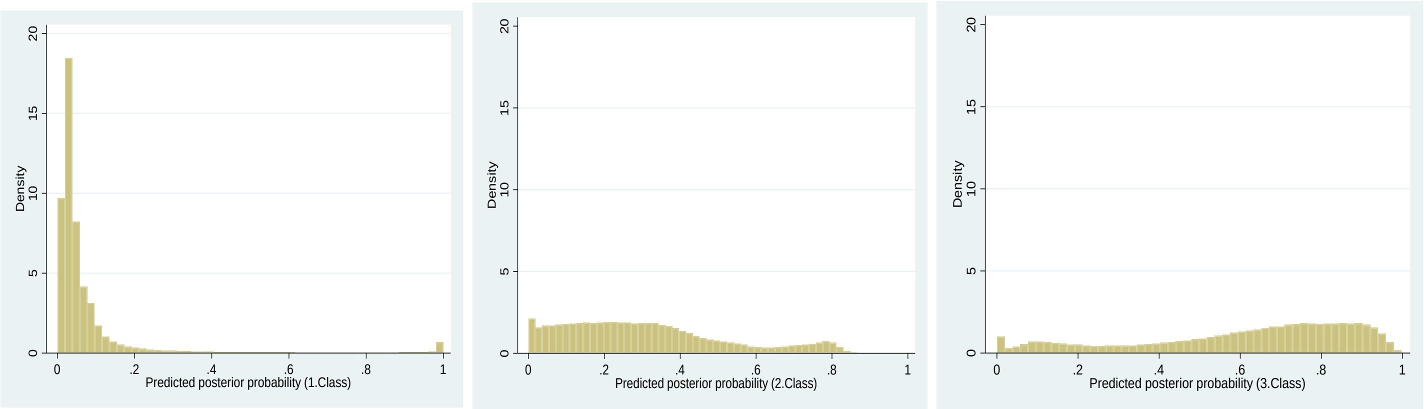
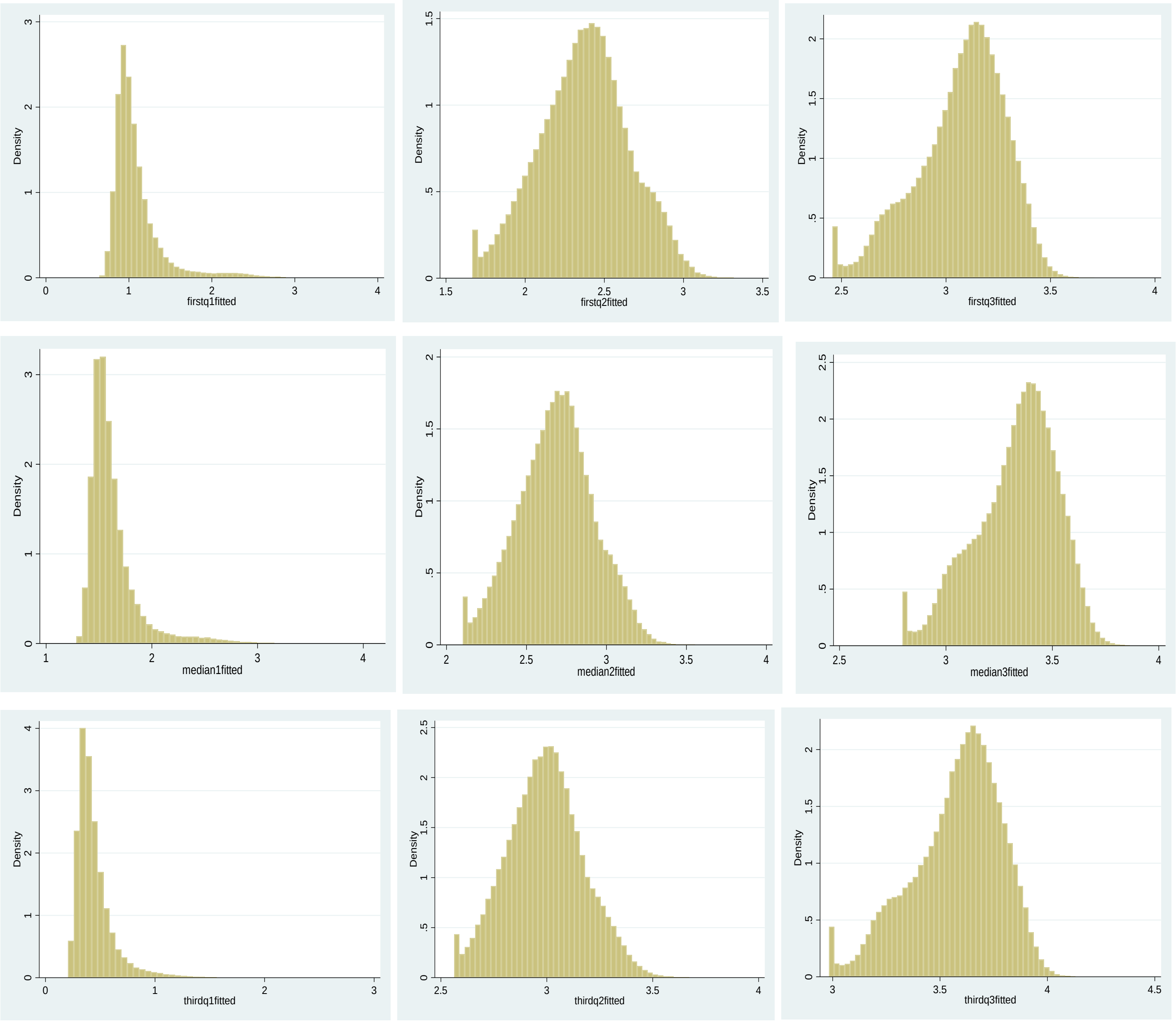


Figure 2: Empirical distributions within each group at the quartiles

Group 1 | Group 2 | Group 3



τ=.25

τ=.5

τ=.75